Analiza krajšanja slovenskih sporočil na družbenem omrežju Twitter

Teja Goli[1], Eneja Osrajnik[2], Darja Fišer[3]

[1]Kropa
[2]Maribor
[3]Univerza v Ljubljani

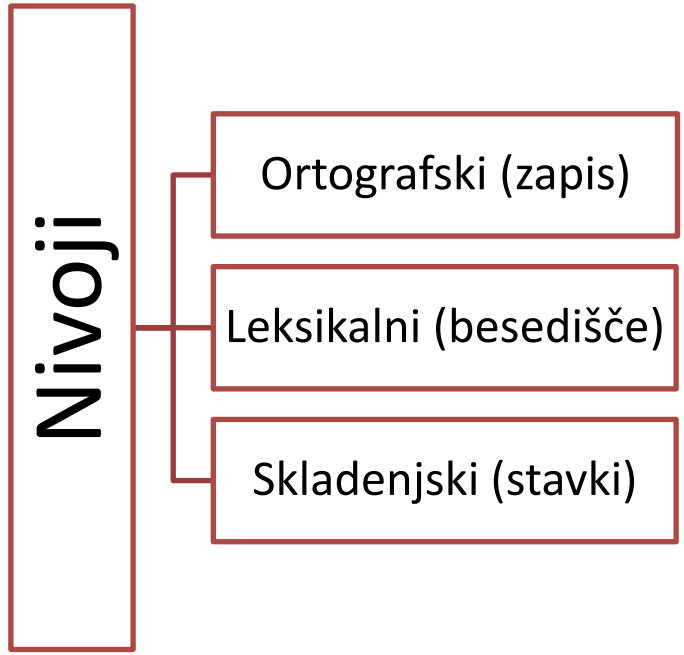
Ljubljana, September 2016

Pregled

- Zasnova raziskave
  - Predmet analize
  - Določanje tipologije
  - Potek označevanja
- Analiza
- Sklep
- Možnosti nadaljnjih raziskav

Zasnova raziskave

- PREDMET ANALIZE
  - Na kakšne načine slovenski uporabniki Twitterja krajšajo svoja sporočila?
  - Kot krajšave smo upoštevali vse pojave, pri katerih izbrana oblika obsega manj znakov kot polno razvezana beseda oz. besedna zveza.
- DOLOČANJE TIPOLOGIJE
  - Poimenovanje nivojev za razvoj tipologije
  - Najmanj dve in največ štiri podkategorije



1 Ortografski nivo (raven zapisa)

1.1 Izpuščanje (črk)	1.2 Opuščanje (nečrkovnih fenomenov)	1.3 Nadomeščanje	1.4 Drugo (**OX**)
1.1.1 na začetku besede	1.2.1 presledkov	1.3.1 večjih nizov črk z manjšim (**ONR**)	
1.1.1.1 samogl. (**OIZV**)	1.2.1.1 pri ločilih (**OOPP**)	1.3.2 s številčnimi homofoni (**ONŠ**)	
1.1.1.2 soglas. (**OIZK**)	1.2.1.2 med besedami (**OOPB**)	1.3.3 domačih črk s tujejezičnimi (**ONT**)	
1.1.1.3 obojega (**OIZO**)	1.2.2 ločil	1.3.4 tujih črk z domačimi črkami (**OND**)	
1.1.2 na sredi besede	1.2.2.1 končnega ločila (**OOLK**)	1.3.5 z logogrami (**ONL**)	
1.1.2.1 samogl. (**OISV**)	1.2.2.2 pike za okrajšavami (**OOLO**)	1.3.6 z emotikoni/emojoji/piktogrami (**ONE**)	
1.1.2.2 soglas. (**OISK**)	1.2.2.3 delov večdelnih ločil (**OOLV**)		
1.1.2.3 obojega (**OISO**)			
1.1.3 na koncu besede			
1.1.3.1 samogl. (**OIKV**)			
1.1.3.2 soglas. (**OIKK**)			
1.1.3.3 obojega (**OIKO**)			

1 Leksikalni nivo		
1.1 Ustaljene krajšave	1.2 Neustaljene krajšave	1.3 Drugo (**LX**)
1.1.1 začetnice/akronimi (**LUZ**)	1.2.1 žanrsko-specifične (**LNŽ**)	
1.1.2 okrajšave (**LUO**)	1.2.2 krajšani neologizmi (**LNN**)	
1.1.3 simboli/formule/krnjene besede (**LUS**)	1.2.3 individualne (**LNI**)	

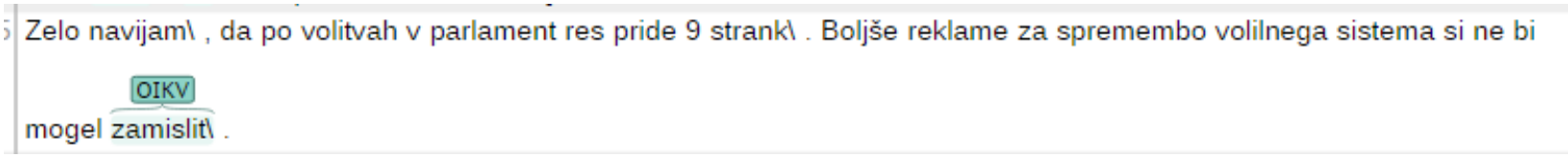
1 Skladenjski nivo
1.1 Izpust glagola
1.1.1 pomožnega (**SGP**)
1.1.2 Izpust glavnega glagola (**SGG**)
1.2 Izpust zaimka (**SZ**)
1.3 Izpust predloga (**SD**)
1.4 Izpust samostalnika (**SS**)
1.5 Drugo (**SX**)

Potek označevanja

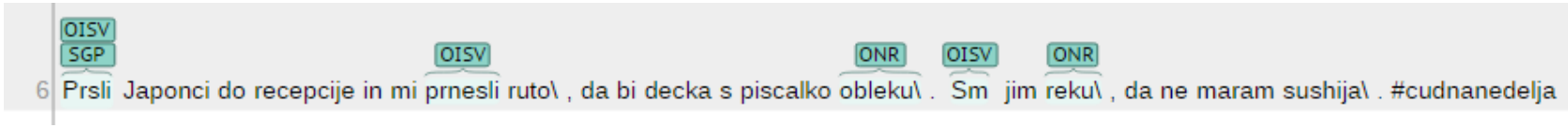
- 3 faze označevanja (2 označevalki) – orodje WebAnno
  - 50 naključno izbranih testnih tvitov za razvoj tipologije
  - Dvojno označevanje 100 jezikovno standardnih in tehnično nestandardnih tvitov + 100 jezikovno in tehnično nestandardnih tvitov
  - Kuriranje oznak → diskrepance (zaradi različnega vrstnega reda oznak = nerealna nesoglasja)
- 600 enojno označenih tvitov iz preostalih kategorij (ne)standardnosti (več o tem: Nikola Ljubešić et al. *Predicting the level of text standardness in user-generated content*. 2015.)
- Problem: narečne besede → težko ločevati med rabo nestandardne leksike in krajšanjem.
  - Nestandardnih besed (*šipa*, *ketna*, *čuza* ipd.) nismo dojemali ko pojave krajšanja, čeprav so krajše od svojih standardnih ustreznic, medtem ko smo kot krajšanje razumeli npr. nestandardne redukcije končnic besed (*jedu* namesto *jedel*).

Vrsta tvitov glede na stopnjo standardnosti	Število tvitov
tehnično in jezikovno standardni (T1L1)	200
tehnično standardni, jezikovno nestandardni (T1L3)	200
tehnično nestandardni, jezikovno standardni (T3L1)	200
tehnično in jezikovno nestandardni (T3L3)	200

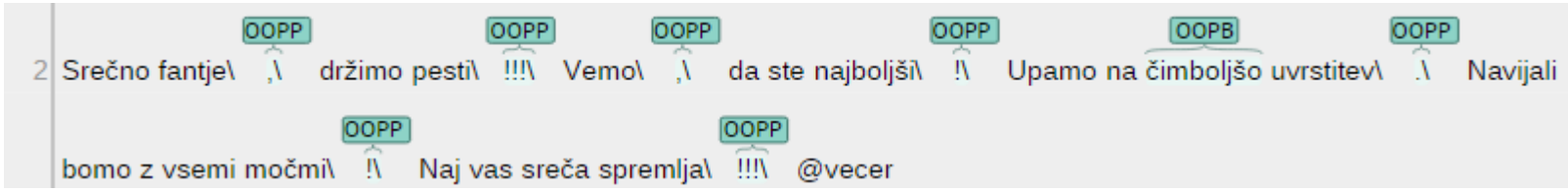
- T1L1:



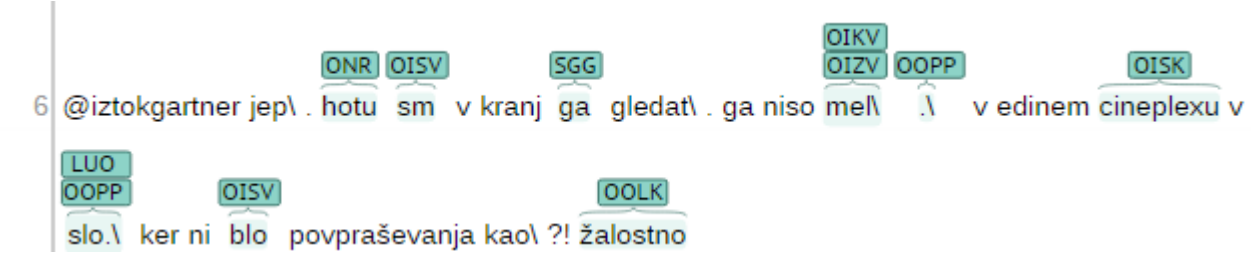
- T1L3:



- T3L1:



- T3L3:



Ugotovitve

- Pričakovano največ redukcij pri tehnično in jezikovno nestandardnih tvitih (43 %).
- 800 tvitov, 3464 pojavov krajšanja:
  - Največ na nivoju zapisa (87 %)
  - Dobrih 700 pojavnic z vsaj dvema oznakama
  - 21 pojavnic s tremi
  - 2 pojavnici s štirimi
- 32 različnih (pod)kategorij
- 3 predvidene kategorije niso prišle v poštev (nadomeščanje s številčnimi homofoni, nadomeščanje z logogrami, nadomeščanje z emotikoni, emojiji, piktogrami)
- 21 izločenih tvitov (avtomatsko generirani, tujejezični, tvit v bohoričici)

Nivo	Stopnja stand.	Št.	%
Zapis (86,92 %)	T1L1	185	5,34 %
	T1L3	781	22,55 %
	T3L1	716	20,67 %
	T3L3	1329	38,37 %
Leksika (11,61 %)	T1L1	85	2,45 %
	T1L3	114	3,29 %
	T3L1	74	2,14 %
	T3L3	129	3,72 %
Skladnja (1,47 %)	T1L1	12	0,35 %
	T1L3	12	0,35 %
	T3L1	17	0,49 %
	T3L3	10	0,29 %

```
$0
1 @murekar Poslediza okamenelega terga nepremizhnin\ , ne vsrok\ . Ştarzi bodo prodali\ /\ dali v\ '\ najem\ , ko drugi ne bodo plazhevali sa njihove hishe\ .
```

Raven zapisa

- Najbolj raznolika raven
- Največ krajšav v tej kategoriji
- Ločila > črke
- Opuščanje presledkov pri ločilih (30 %)
- Izpuščanje samostalnikov na koncu besede (18 %)
- Opuščanje končnega ločila (14 %)
- Izpuščanje samostalnika na sredi besede (14 %)
- Nadomeščanje daljših nizov črk s krajšimi (7 %)

Nivo1	Nivo2	Nivo3	Primer	Pogostost
Izpušč.	začetek b.	samostalnik	mam	1,89 %
		soglasnik	lej	0,20 %
		oboje	koj	0,07 %
	sredina b.	samostalnik	bedn	14,31 %
		soglasnik	današnega	1,13 %
		oboje	kera	1,13 %
	konec b.	samostalnik	anglesk	17,77 %
		soglasnik	ka	1,13 %
		oboje	lah	1,93 %
Opuš.	presl.	pri ločilih	prepričano,da	29,76 %
		med besedami	inče tood njihni	3,49 %
	ločil	konč.	Ti to iz rokava streseš	14,48 %
		pike za okr.	Slo	1,43 %
		večbes. ločila	Kajmak in marmelada..	3,42 %
Nadom.	daljših nizov s krajšimi		ponuju	6,91 %
	s številčnimi homofoni		-	0,00 %
	domačih črk s tujejezičnimi		explozij	0,53 %
	tujih črk z domačimi		Tviter	0,37 %
	z logogrami		-	0,00 %
	z emotikoni		-	0,00 %
drugo			EPPja, 13incni	0,46 %

Raven zapisa

- Zanimivosti:
  - Nadomeščanje domačih črk s tujejezičnimi: 9 od 11 primerov „x“ namesto „ks“ (*expert*); *ql*, *vö* (=ven)
  - Nadomeščanje tujejezičnih črk z domačimi: *pusiji*, *tviter*, *stori* (beseda v resnici ni bila krajša zaradi nadomeščanja)
  - Nismo identificirali nadomeščanja s številčnimi homofoni, logogrami, emojiji …

Leksikalna raven – ustaljene k.

- Najpogostejša je raba ustaljenih začetnic/kratic (35 %), od skupno 140 so kar 104 različne (*EU*, *ZDA*, *PS*, *SDS* …)
- Simboli: najpogostejša raba znaka +, ki največkrat nadomešča veznik *in*, ter raba črke *x*, ki predstavlja „*krat*“ (15 %)
- Ustaljene okrajšave (raba pike ni vplivala na določitev njihove (ne)ustaljenosti, saj smo manjkajoča ločila označili na ortografskem nivoju). (1 2%)
  - cca, dr., oz., itd.

Nivo1	Nivo2	Primer	Pogostost
ustalj. krajšave	začetnice/ kratice	BDP	34,83 %
	okrajšave	št., slo.	11,69 %
	simboli, formule, krnjene besede	€, +	14,93 %
neustal. krajšave	žanrsko-specifične	Tw.	12,44 %
	krajšani neologizmi	appi, Zoki	9,45 %
	priložnostne	sod[nik]	12,94 %
drugo		e-sožalje	3,73 %

Leksikalna raven: neustaljene k.

- Neustaljene krajšave so ene izmed najzanimivejših: izrazit strateški značaj časovnega, prostorskega oz. tehničnega krajšanja + evidentna kreativnost uporabnikov.

- Priložnostne krajšave = ad hoc krajšanja, pri katerih avtor uporabi skovanko, ki jo lahko razumemo zgolj s pomočjo konteksta, v katerem se pojavi.
  - PV → predsednik vlade
  - KlinecTM → Univerzitetni klinični center
  - upravič → upravičiti
  - odl → odločitev
- Neustaljene žanrske krajšave = vse tiste krajšave, ki se izgovarjajo po posameznih črkah in so med drugim tipične za elektronska besedila.
  - rt, btw, fb, lj, jbt, Tw

- Neologizmi = tiste krajšave, ki se za razliko od žanrskih izgovarjajo kot besede. Sem smo umestili tudi vzdevke.
  - alko, app, simultanka, Zoki, Bojči

Skladenjska raven

- Problematično določanje skladenjskih izpustov, zaradi otežene distinkcije med namernim krajšanjem in elipsami, ki pripomorejo k večji koherentnosti besedila.

- Zato smo označili le tiste izpuste besednih vrst, ki niso bili posledica „običajnega“ izpusta zaradi sobesedila.

  - Včeraj [**SGG**]{sem} bil na kuhančku v Ljubljani.

- Največkrat (61 %) je šlo za izpust pomožnega glagola (=najmanjši vpliv na razumevanje besedila).

Nivo1	Nivo2	Primer	Pogostost
izpust glag.	pomož.	uf, zdej mi ze vec stvari jasnih hehe thx za info /…/	60,78 %
	glavnega	Bi blo treba tiralico?	13,73 %
izpust zaimka		…strinjam popolnoma…	3,92 %
izpust predloga		ocene /…/ se izkažejo predvsem [SD] politično motivirane	3,92 %
izpust samostalnika		/…/ uhhh ova je lejpa ka ma na boki [SS] od vseh knjig…	7,84 %
drugo		/…/ kjer lahko lajkaš to stran, ne pa vas morm dodat za frenda /…/	9,80 %

Kategorija ***Drugo***

- Kategorija **Drugo** na vseh nivojih za označevanje nepredvidenih elementov krajšanja:
  - Zapis: manjkajoč vezaj (*euprava*, *90ta*, *rtvja*)
  - Leksika: izpeljanke iz ustaljenih kratic (*Desusovec*), beseda *ex* (*ex politični zapornik*)
  - Skladnja: odrezani tviti oz. tviti, razdeljeni na več delov
    - punce :) vasja danc na bazenu v sgjo od 3 ure dalje :) tk da lahka pridete ko ma tako zeljo vas s [**SX**] http://t.co/qvk6jeMeBc

Sklep

- Zabeležili smo 3464 pojavov krajšanja na 800 tvitih. → Le dobrih 10 % zapisov ni vsebovalo nobenega krajšanja. → Slovenski tviteraši so zelo naklonjeni krajšanju besedil!

- Največ krajšanj v tehnično in jezikovno nestandardnih tvitih.

- Največ na nivoju zapisa, prednjači opuščanje presledkov ob ločilih ter izpuščanje samoglasnikov na koncu besed. Krajšanje na skladenjski ravni je redko (manj kot 2 %) → razumljivost besedila je pomembna.

- Aspekt kreativnosti uporabnikov: priložnostne in žanrsko-specifične krajšave.

- Motivacija za redukcije = ???
  Razlogi za izpuste niso enoznačni in težko določimo pravo motivacijo za krajšanje

  – Tehnične bližnjice z izpuščanjem presledkov in večdelnih ločil.
  – Zavestna odločitev uporabnikov → prostorska omejitev, kreativnost, kriptiranje, osebni slog, napake …

Nadaljne raziskave

- Raziskavo lahko dopolnili z analizo drugih žanrov v korpusu Janes ter primerjali pojave krajšanja slovenskih uporabnikov, kadar tvitajo v različnih jezikih.
- Primerjava krajšanja v dolgih in kratkih tvitih. → Vpliv tehničnih okoliščin komuniciranja.
- Sporočila na mobilnih napravah vs. na računalnikih.
- Opazovanje uporabe krajšav v različnih časovnih obdobjih. → Ali krajšanje narašča, upada ali stagnira + ali postaja bolj/manj homogeno?
- Daljšanje v zapisu? Usklajenost z glasovno realizacijo besed pri prevzemanju iz jezikov z veččrkji (q – ku, x – ks, nasproti ch – č).

Literatura

- Christian M. Alis in May T. Lim. 2013. Spatio-Temporal Variation of Conversational Utterances on Twitter http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0077793#s3.
- Jannis Androutsopoulos in Gurly Schmidt. 2001. SMS-Kommunikation: Ethnografische Gattungsanalyse
- am Beispiel einer Kleingruppe. V: Meer, D., ur., Zeitschrift für Angewandte Linguistik. Bd. 36, Frankfurt/Main, 49–80.
- Markus Bieswanger. 2007. abbrevi8 or not 2 abbrevi8: A contrastive analysis of different space and time- saving strategies in English and German text messages. V: Texas Linguistics Forum, volume 50.
- David Crystal. 2001. Language and the Internet. Cambridge: Cambridge University Press.
- Jaka Čibej, Darja Fišer in Tomaž Erjavec. (V tisku). Normalisation, Tokenisation and Sentence Segmentation of Slovene Tweets.
- Nicola Döring. 2002. Kurzm. wird gesendet – Abkürzungen und Akronyme in der SMS-Kommunikation. V: Muttersprache - Vierteljahresschrift für deutsche Sprache, 112 (2), 97–114.
- Richard Eckart de Castilho et al. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN V: Proceedings of the CLARIN Annual Conference (CAC) 2014. CLARIN ERIC, October 2014. https://www.clarin.eu/content/papers-posters-and-demos-cac2014.
- Stephan Gouws et al. 2011. Contextual Bearing on Linguistic Variation in Social Media. V: Proceedings of the Workshop on Language in Social Media (LSM 2011). Portland, Oregon. 20–29. http://dl.acm.org/citation.cfm?id=2021113
- Ylva Hård af Segerstad. 2002. Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication. Göteborg: Göteborg University.
- Susan C. Herring. 2001. Computer-mediated discourse. V: D. Schiffrin, D. Tannen, & H. Hamilton, ur., (pp.The Handbook of Discourse Analysis 612-634). Oxford: Blackwell.
- Urška Jarnovič. 2007. Diskurzivne značilnosti SMS-ov. Jezik in slovstvo, 52 (2): 61–79. Ljubljana: Slavistično društvo Slovenije.
- Mojca Kompara. 2009. Prepoznavanje krajšav v besedilih. V: Peter Weiss, ur., Jezikoslovni zapiski 15, št. 1–2. 95–112. Založba ZRC, Ljubljana.
- Nikola Ljubešić, Darja Fišer, Tomaž Erjavec, Jaka Čibej, Dafne Marko, Senja Pollak in Iza Škrjanec. 2015. Predicting the level of text standardness in user-generated content. Zbornik konference RANLP 2015, 7.–9. september 2015, str. 371–378, Hissar, Bolgarija.
- Mija Michelizza. 2012. Ločila in druga pisna znamenja v elektronskih besedilih. V: Nataša Jakop, Helena Dobrovoljc, ur., Pravopisna stikanja: razprave o pravopisnih vprašanjih. 151–162. Založba ZRC, Ljubljana.
- Richard Rogers. 2014. Debanalising Twitter. The transformation of an object of Study. V: Katrin Weller et al., ur., Twitter and Society. IX–XXVI. Peter Lang Publishing, Inc., New York.
- Peter Schlobinski et al. 2001. Simsen. Eine Pilotstudie zu sprachlichen und kommunikativen Aspekten der SMS-Kommunikation. Networx 22. Retrieved July 1, 2006, from http://www.mediensprache.net/networx/networx-22.pdf.
- Malin Sveningsson. 2001. Creating a Sense of Community: Experiences from a Swedish Web Chat. The TEMA Institute, Dept. of Communication Studies. Linköping, Linköping University: 250.
- Crispin Thurlow. 2003. Generation Txt? The sociolinguistics of young people's text- messaging. Discourse Analysis Online, 1 (1).